The Detection Dividend: Staffing Mandates and Endogenous Regulatory Metrics in U.S. Nursing Homes*

APEP Autonomous Research

April 2, 2026

Abstract

When policy changes the intensity of regulatory monitoring, do administrative metrics still measure what they were designed to measure? Exploiting staggered adoption of nursing home staffing mandates across six U.S. states, I show that mandates increase detected deficiency citations by 43% while simultaneously *improving* infection control outcomes. The additional citations concentrate in observation-dependent categories, shift toward low-severity findings, and do not appear in complaint-driven deficiencies—which bypass surveyor detection entirely. I call this pattern the *detection dividend*: more staff on the floor during inspections expands the regulatory surface area available to surveyors, generating citations that reflect enhanced observability rather than deteriorating care. These findings demonstrate that administrative compliance metrics are endogenous to the policies they are used to evaluate, with implications for quality rating systems, pay-for-performance, and regulatory design beyond healthcare.

JEL Codes: I11, I18, J23, K32, L51

Keywords: endogenous regulatory metrics, detection, nursing homes, staffing mandates, deficiency citations, Five-Star ratings

*This paper is a revision of APEP-0959. See https://github.com/SocialCatalystLab/ape-papers/tree/main/apep_0959_v1. Autonomous Policy Evaluation Project. Correspondence: scl@econ.uzh.ch

1. Introduction

What does a regulatory metric measure when the policy it evaluates also changes the technology of measurement? This question arises whenever a government intervention alters not only the behavior being regulated but also the probability that violations of the regulation are detected. In such settings, observed compliance statistics become endogenous to the policy itself—a complication that standard program evaluation typically ignores.

This paper documents a concrete and consequential instance of this problem. U.S. nursing home staffing mandates—laws requiring minimum hours of direct nursing care per resident per day—are associated with increases in detected deficiency citations of approximately 1.2 per inspection survey in New York (the primary specification) and 2.1 per survey in the pooled six-state sample (43% of the control mean). Yet the same mandates *reduce* infection control deficiencies, one of the most directly staffing-sensitive citation categories, by 0.03 per survey ($p < 0.01$). Report-dependent deficiencies, which originate from resident and family complaints rather than surveyor observation, show no change whatsoever. The increase concentrates in observation-dependent citation categories—those requiring a surveyor to directly witness staff-resident interactions—and shifts toward low-severity administrative findings rather than citations involving actual patient harm.

I interpret this pattern as a *detection dividend*. More nurses on the floor during unannounced inspections means more care interactions for surveyors to observe, more documentation to review, and more staff to interview. The regulatory surface area expands. Surveyors, following their standardized protocols, discover more minor deviations from the approximately 180 federal requirements—not because care has worsened, but because the inspection technology has become more thorough. The facility with twelve aides on the morning shift presents twelve sets of hand-hygiene practices, medication-pass procedures, and resident-interaction opportunities for a surveyor to evaluate. The facility with eight presents eight. The deficiency count reflects both care quality and observational opportunity, and staffing mandates change both simultaneously.

The detection dividend matters for three reasons, each extending beyond the nursing home sector. First, deficiency citations are the primary input to the CMS Five-Star Quality Rating System, which shapes consumer choice, Medicaid reimbursement adjustments, and regulatory scrutiny for over 15,000 facilities nationwide (Werner and Dudley, 2012; Konetzka et al., 2021). If mandates mechanically increase citations through enhanced detection rather than quality deterioration, the rating system penalizes precisely the facilities that comply with staffing requirements—a perverse feedback loop in which doing what the regulator demands worsens the regulator's own assessment of performance.

Second, the finding speaks directly to the political economy of the 2024 federal minimum staffing rule, the first in Medicare's history, which was finalized in March 2024 and suspended by Congress within a year (Centers for Medicare & Medicaid Services, 2024, 2025). Opponents cited deficiency trends in mandate states as evidence that staffing floors do not improve quality. My results suggest this evidence is misleading: the raw data confound detection intensity with care quality, and disentangling the two requires the kind of decomposition I perform here.

Third, the detection dividend is an instance of a general phenomenon in regulatory economics. Whenever enforcement effort is endogenous to the policy being enforced, measured compliance becomes unreliable as a welfare indicator. Duflo et al. (2013) document this in Indian pollution auditing: under the status quo, third-party auditors reported only 7% of plants as violators when 59% actually exceeded standards. Chalfin and McCrary (2018) show that police staffing affects crime measurement through reporting channels. Olken (2007) demonstrates that increasing audit probability from 4% to 100% reduces measured corruption by 8 percentage points in Indonesian village road projects. Christensen et al. (2013) find that capital-market benefits attributed to IFRS adoption actually reflected concurrent enforcement changes. In each case, the observed outcome is a joint product of the underlying behavior and the detection technology, and policy changes both. The present paper adds a new setting—healthcare regulation—and a distinctive mechanism: the "third party" whose presence changes detection intensity is the regulated entity's own mandated workforce.

The empirical strategy exploits staggered adoption of quantitative hours-per-resident-per-day (HPRD) staffing floors across six states—Connecticut, Rhode Island, California, Arizona, Washington, and New York—between 2017 and 2022. I use CMS Health Deficiency data covering 418,972 citation records across 14,636 facilities, merged with provider characteristics and staffing information from the Payroll-Based Journal system. The primary specification focuses on New York's 2022 Safe Staffing for Quality Care Act, which provides five pre-treatment years. Both the NY and pooled specifications show a significant $t-4$ pre-trend that I discuss transparently; the $t-3$ and $t-2$ coefficients are clean in both. The detection-mode *pattern*—not any single aggregate estimate—is the paper's main evidence.

The core empirical architecture rests on a detection-sensitivity taxonomy. I classify all deficiency categories into three groups based on the CMS State Operations Manual's inspection methodology: *observation-dependent* violations (discovered when surveyors directly witness care interactions), *documentation-dependent* violations (discovered through record and chart review), and *report-dependent* violations (initiated by resident or family complaints and investigated independently of routine surveys). The detection dividend predicts a specific sign pattern across these categories. If mandates genuinely deteriorated care quality,

all three categories should increase. If mandates change only the detection technology, observation-dependent and documentation-dependent citations should rise while report-dependent citations—which bypass the surveyor observation channel entirely—should remain flat. The data match the detection prediction precisely: observation-dependent deficiencies increase by 1.294 per survey ($p < 0.01$) and documentation-dependent by 0.205 ($p < 0.10$) in New York; report-dependent deficiencies show a point estimate of -0.310 ($p = 0.23$), indistinguishable from zero.

I complement this decomposition with a severity analysis. The CMS scope-severity grid classifies each citation on a scale from A (isolated, no actual harm, minimal potential) through L (widespread, immediate jeopardy). If the detection dividend reflects enhanced observability of minor regulatory deviations, the additional citations should concentrate at the low end of the severity distribution. They do: in the pooled specification, moderate-severity citations (grades D through F) increase by 1.958 per survey ($p < 0.05$), while citations involving actual patient harm (grades G through I) increase by only 0.057 ($p < 0.01$, but economically trivial relative to the total effect), and citations at the jeopardy level (grades J through L) show no change. In the New York specification, high-severity citations (grades G through L) have a point estimate of -0.002—precisely zero.

Identification faces two principal challenges that I address but do not fully resolve. First, with only six treated states, state-clustered inference is inherently limited. I report state-clustered standard errors as the primary inference (with facility-clustered SEs as a robustness check), leave-one-state-out sensitivity, and HonestDiD bounds under the Rambachan and Roth (2023) relative magnitudes framework. At exact parallel trends ($\bar{M} = 0$), the 95% robust confidence interval for the pooled effect is $[-0.91, 1.08]$, which includes zero—the data cannot reject a null effect even under the most favorable assumptions. By $\bar{M} = 1$, the interval widens to $[-6.06, 6.16]$. This sobering result underscores why the paper relies on the *pattern* across detection modes rather than any single aggregate point estimate. Second, the pooled Sun-Abraham event study shows a concerning pre-trend coefficient at $t-4$ ($+2.23$, $p < 0.01$), though $t-2$ is small and insignificant ($+0.05$). The $t-3$ coefficient ($+0.93$) is marginally significant. The New York event study, which is the paper's primary specification, shares the $t-4$ anomaly ($+2.887$, $p < 0.001$) but has clean $t-3$ ($+0.569$, not significant) and $t-2$ (-0.266, not significant) coefficients. The post-treatment effect builds gradually over $t+2$ through $t+3$, consistent with the 12-month inspection cycle. Neither specification provides a definitive causal estimate, but both display the same mechanism pattern across detection modes.

Heterogeneity analysis reveals that for-profit facilities—which constitute 79% of the sample and historically maintain lower staffing ratios—drive the detection dividend, with an increase

of 2.361 per survey compared to 0.796 for nonprofits. This is consistent with the mechanism: for-profit facilities, starting from lower staffing baselines, experience the largest marginal increase in regulatory surface area when mandates bind.

This paper contributes to three literatures. First, it contributes to the economics of nursing home regulation, where a large body of work examines the relationship between staffing and quality (Harrington et al., 2000; Castle and Ferguson, 2011; Lin, 2014), the effects of staffing mandates on staffing levels and input substitution (Bowblis, 2011; Matsudaira, 2014; Bowblis, 2013; Werner et al., 2026), and the consequences for market structure and exit (Bowblis and Ghattas, 2017). I add the first evidence that staffing mandates change the informational content of the regulatory metrics used to evaluate them. Second, I contribute to the economics of enforcement and monitoring. The theoretical literature—Becker (1968), Stigler (1970), Polinsky and Shavell (2000)—models detection probability as a policy instrument; I show that detection probability can also be an unintended byproduct of a policy aimed at something else entirely. The empirical enforcement literature—Duflo et al. (2013), Olken (2007), Shimshack and Ward (2005), Gray and Shimshack (2011)—treats changes in measured violations as evidence of enforcement effectiveness; I demonstrate that in some settings, more measured violations reflect enhanced measurement, not changed behavior. Third, I contribute to the literature on performance metrics and accountability gaming (Dranove et al., 2003; Jacob, 2005; Neal and Schanzenbach, 2010), but with a crucial distinction: the detection dividend is not strategic manipulation by the regulated entity. Nursing homes do not game inspections by hiring more staff. The measurement distortion arises mechanically from the coupling between the policy instrument (staffing floors) and the measurement technology (surveyor observation of staff-resident interactions).

The remainder of the paper proceeds as follows. Section 2 presents a simple conceptual framework that formalizes the detection dividend and derives testable predictions. Section 3 describes the institutional setting. Section 4 introduces the data and empirical strategy. Section 5 presents the main results. Section 6 examines identification and robustness. Section 7 discusses implications and limitations. Section 8 concludes.

2. Conceptual Framework

This section develops a simple framework to organize the paper's predictions and interpretation. The framework is intentionally reduced-form—it is a taxonomy of channels, not a structural model. Its purpose is to make precise the distinction between changes in true violations and changes in detection intensity, and to derive observable implications that can be tested in the data.

2.1 Observed Violations as a Joint Product

Let V_i^* denote the true number of regulatory violations at facility i—the violations that would be found under perfect, exhaustive inspection. Let $D_i \in [0, 1]$ denote the *detection rate*: the probability that a given true violation is discovered during a standard survey. The number of observed deficiency citations is:

$$V_i = D_i \cdot V_i^* \tag{1}$$

A policy P (here, a staffing mandate) may affect both components. The total derivative of observed violations with respect to the policy is:

$$\frac{dV_i}{dP} = \underbrace{\frac{\partial D_i}{\partial P} \cdot V_i^*}_{\text{Detection channel}} + \underbrace{D_i \cdot \frac{\partial V_i^*}{\partial P}}_{\text{Quality channel}} \tag{2}$$

The first term captures the detection dividend: even if true violations are unchanged (or decline), observed violations can increase if the policy raises the detection rate. The second term captures the standard program-evaluation object of interest: the effect of the policy on actual care quality.

2.2 Why Staffing Mandates Affect Detection

In nursing home inspections, the detection rate D_i depends on the *regulatory surface area* available to surveyors: the number of staff-resident interactions that can be observed, the volume of documentation that can be reviewed, and the number of personnel who can be interviewed. A staffing mandate that raises the number of nurses present during an inspection mechanically expands this surface area. If a surveyor observes ten aide-to-resident interactions during a medication pass, the probability that at least one deviates from the 180-item regulatory checklist is higher than if the surveyor observes six.

Formally, let n_i denote the number of staff present during the survey. If each staff member independently generates regulatory exposure with some probability δ, then the expected number of detectable violations is increasing in n_i:

$$\mathbb{E}[V_i \mid n_i] = n_i \cdot \delta \cdot q_i + g(n_i) \tag{3}$$

where q_i is a facility-specific quality parameter (lower q_i means better care) and $g(n_i)$ captures the documentation and interview channel (more staff $\rightarrow$ more records to review). A staffing mandate raises n_i, which increases $\mathbb{E}[V_i]$ through both the observation channel ($n_i \cdot \delta \cdot q_i$) and

the documentation channel ($g(n_i)$), even if q_i remains constant or improves.

2.3 Testable Predictions

The framework generates four predictions that distinguish the detection dividend from a pure quality deterioration story:

Prediction 1 (Detection-mode asymmetry). If the detection channel dominates, citation increases should concentrate in observation-dependent and documentation-dependent categories—those where more staff mechanically expands the regulatory surface area. Report-dependent deficiencies, which are initiated by resident and family complaints and investigated independently of routine surveys, should not respond to staffing changes.

Prediction 2 (Severity composition). If the additional citations reflect marginal regulatory deviations that become detectable only with enhanced observation, they should concentrate at the low end of the severity distribution (grades A through F: no actual harm). Citations involving actual patient harm (grades G through L) should not increase—and may decrease if staffing genuinely improves care.

Prediction 3 (Countervailing quality signal). At least one clinical quality outcome that is directly staffing-sensitive—such as infection control, which depends on hand hygiene, isolation protocols, and staffing adequacy—should improve, providing a positive signal that coexists with the negative signal from total deficiency counts.

Prediction 4 (Heterogeneity by baseline staffing). The detection dividend should be largest at facilities with the lowest baseline staffing, where the mandate-induced increase in regulatory surface area is greatest. For-profit facilities, which historically maintain lower staffing ratios (Harrington et al., 2012), should show larger effects than nonprofits.

Section 5 tests each of these predictions.

3. Institutional Background

3.1 Nursing Home Inspection and Deficiency Citations

The Omnibus Budget Reconciliation Act of 1987 (OBRA-87) established the federal framework for nursing home quality regulation. Medicare- and Medicaid-certified facilities must undergo unannounced health inspections at least every 15 months, with a statewide average target of 12 months. State survey agencies conduct these inspections, evaluating compliance across approximately 180 regulatory requirements—called "tags"—covering resident rights, quality of care, quality of life, infection control, nutrition, pharmacy services, and the physical environment (Centers for Medicare & Medicaid Services, 2023).

Each identified violation is documented as a deficiency citation. The citation records the specific regulatory tag, a description of the violation, and a scope-severity classification on a 12-level grid. Scope ranges from isolated (affecting one or a few residents) to pattern (affecting multiple residents) to widespread (affecting most residents). Severity ranges from minimal potential for harm (level 1) to actual harm (level 3) to immediate jeopardy (level 4). The resulting letter grades—A through L—determine enforcement consequences: levels A through D typically require a voluntary plan of correction; levels E through H may trigger civil monetary penalties or denial of payment for new admissions; levels I through L (immediate jeopardy) can trigger decertification.

A critical feature for this paper is that the standard health survey is *observational*. Surveyors spend several days in the facility, directly observing care delivery, interviewing staff and residents, and reviewing documentation. The number and quality of staff present during the survey directly affect the scope of what surveyors can observe. This creates the mechanical link between staffing and detection that underlies the detection dividend.

3.2 State Staffing Mandates

While OBRA-87 requires "sufficient" nursing staff, it does not specify a quantitative floor. Beginning in the late 1990s, states began enacting minimum staffing requirements expressed as hours per resident per day (HPRD). As of 2025, approximately 35 states have some form of staffing standard, but only a subset specify quantitative HPRD floors rather than qualitative staffing-plan requirements (Harrington et al., 2020).

My treatment group consists of six states that enacted or substantially updated quantitative HPRD floors during the Payroll-Based Journal data window (2017–2026): Connecticut and Rhode Island (2017), California (2018, raising its floor to 3.5 total HPRD under AB 2079), Arizona and Washington (2019), and New York (2022, the Safe Staffing for Quality Care Act requiring 3.5 total HPRD with a 2.2 CNA floor). Six additional states with mandates predating the data window—Florida, Illinois, Arkansas, Oregon, Pennsylvania, and Massachusetts—are excluded as always-treated units. An important caveat: Connecticut and Rhode Island, treated in 2017, have essentially no pre-treatment observations in the data window. Their inclusion in the pooled specification adds cross-sectional variation but no pre-trend information. The Sun-Abraham estimator handles this by using not-yet-treated units as controls, but these two cohorts contribute little to identification.

The primary specification centers on New York's 2022 mandate: a single treatment cohort with five pre-treatment years (2017–2021) and three full post-treatment years (2023–2025), plus a partial 2022 (the mandate took effect in January but many facilities were inspected before implementation was complete) and partial 2026 (through March). The event study

window spans $t = -4$ to $t = +4$ relative to 2022. The pooled specification uses all six staggered cohorts, with the 2017-treated states contributing only post-treatment variation.

3.3 The Five-Star Quality Rating System

CMS publishes facility-level quality ratings on the Nursing Home Compare website (now Care Compare), using a Five-Star system where one star is "much below average" and five stars is "much above average." The overall rating is a weighted composite of three domains: health inspections, staffing, and quality measures. The health inspection domain—which is constructed directly from deficiency citation counts and severity—is the most heavily weighted component and is the domain most directly affected by the detection dividend.

If staffing mandates mechanically increase deficiency citations through enhanced detection, mandate-compliant facilities will receive *lower* health inspection ratings precisely because they hired more staff. This creates a perverse incentive: the regulatory system punishes the behavior it requires, because the measurement technology conflates compliance with poor performance.

4. Data and Empirical Strategy

4.1 Data Sources

I combine four datasets from the March 2026 release of the CMS Provider Data Catalog.

Health Deficiency Citations. The primary outcome data record every deficiency citation from standard health inspections at Medicare- and Medicaid-certified nursing homes. Each record identifies the facility (by CMS Certification Number), survey date, deficiency tag number, tag category, scope, and severity code. The raw dataset contains 418,972 citation records spanning fiscal years 2017 through 2026 across 14,636 unique facilities. I aggregate citations to the facility-survey level.

Provider Information. Facility characteristics come from the CMS Provider Information file: current staffing levels by type (RN, LPN, CNA, total HPRD), Five-Star ratings, certified bed count, average daily census, ownership type (for-profit, nonprofit, government), chain affiliation, and urban/rural classification. I use this cross-sectional file for the first-stage analysis and to construct facility-level controls.

Detection-Sensitivity Taxonomy. The paper's central empirical innovation is a classification of deficiency tags by *detection mode*. Using the CMS State Operations Manual

(Appendix PP), I classify each of the approximately 180 regulatory tags into one of three categories based on how violations are typically discovered during the inspection process:

- **Observation-dependent** (approximately 130 tags): Violations discovered primarily through direct surveyor observation of care delivery, staff-resident interactions, environmental conditions, or dining services. Examples include medication administration errors observed during a medication pass, improper transfer techniques observed during a resident-handling episode, and dietary service violations observed during mealtimes. These tags account for 232,000 of the 418,972 total citations in the data.
- **Documentation-dependent** (approximately 30 tags): Violations discovered primarily through review of medical records, care plans, incident reports, and administrative documentation. Examples include incomplete assessments, missing physician orders, and inadequate care planning. These account for 54,000 citations.
- **Report-dependent** (approximately 20 tags): Violations investigated in response to complaints filed by residents, families, or staff—not discovered during the routine survey. These enter the system through a separate complaint intake process and are investigated on a different timeline. They account for 132,000 citations.

This taxonomy is the key to distinguishing detection from deterioration. If mandates worsen care quality, all three categories should increase. If mandates primarily expand the observational surface area, only observation-dependent and documentation-dependent categories should respond.

4.2 Analysis Panel

I merge deficiency citations to facility characteristics by CMS Certification Number, creating a panel of 72,730 facility-survey observations across 11,946 unique facilities in 47 states. The panel excludes six always-treated states (those with quantitative HPRD mandates predating 2017). Among the six treatment cohorts, Connecticut and Rhode Island (treated in 2017) have no pre-treatment observations in the data window; they contribute only post-treatment variation in the pooled specification. The remaining four cohorts (CA 2018, AZ/WA 2019, NY 2022) have progressively longer pre-treatment periods. The analysis period spans 2017 through 2026.

Table 1 presents summary statistics by treatment status (mandate-state observations in the post-mandate period vs. all other observations). Treated observations report somewhat fewer total deficiencies (4.22 vs. 4.86), consistent with the composition of treatment states.

Table 1: Summary Statistics: Deficiency Citations by Treatment Status

	Treatment (Mandate states, post)	Control (All other obs.)
Total deficiencies	4.22	4.86
	(5.43)	(5.12)
Observation-dependent	2.35	2.66
Documentation-dependent	0.52	0.66
Report-dependent	1.34	1.54
Low severity (A–F)	4.08	4.56
High severity (G–L)	0.14	0.29
Infection control	0.01	0.02
Observations	20,953	51,777

Standard deviations in parentheses. Treatment group: facility-survey observations in mandate states after mandate adoption. Control group: all other observations in non-always-treated states.

Observation-dependent citations are the largest category (mean 2.35–2.66), followed by report-dependent (1.34–1.54) and documentation-dependent (0.52–0.66). High-severity citations (grades G through L) are rare, accounting for only 3–6% of all citations.

4.3 Empirical Strategy

I estimate the effect of staffing mandates on deficiency outcomes using a two-way fixed effects (TWFE) difference-in-differences specification:

$$Y_{ist} = \alpha_i + \gamma_t + \beta \cdot \text{Post}_{st} + \varepsilon_{ist} \tag{4}$$

where Y_{ist} is the deficiency outcome for facility i in state s during survey period t; α_i are facility fixed effects absorbing all time-invariant facility and state characteristics; γ_t are year fixed effects absorbing common temporal shocks including the COVID-19 inspection disruption; and Post_{st} equals one after state s's quantitative HPRD mandate takes effect. Standard errors are clustered at the state level, the unit of treatment assignment.

The identifying assumption is that, absent the mandate, treated and control facilities would have experienced parallel trends in deficiency citations. I assess this using event-study specifications.

For the New York primary specification, I estimate:

$$Y_{ist} = \alpha_i + \gamma_t + \sum_{k \neq -1} \delta_k \cdot \mathbf{1}[\text{NY}]_i \cdot \mathbf{1}[t = 2022 + k] + \varepsilon_{ist} \tag{5}$$

where the δ_k coefficients trace out the treatment effect path relative to the year before mandate adoption ($k = -1$). For the pooled specification with staggered adoption, I use the Sun and Abraham (2021) interaction-weighted estimator, which is robust to treatment-effect heterogeneity across cohorts.

What This Comparison Can and Cannot Identify. The DiD design identifies the causal effect of staffing mandates on deficiency citations under parallel trends. It cannot, however, decompose this effect into detection and quality channels—that decomposition relies on the cross-category predictions of the conceptual framework and is therefore an interpretation aided by theory, not a separate source of causal identification. The complaint-deficiency placebo provides the strongest single piece of evidence for the detection interpretation, because the complaint channel is mechanically unrelated to surveyor observation during routine surveys. But even this test is not definitive: if mandates simultaneously improved care quality (reducing complaints) and increased detection (raising routine citations), the zero complaint effect would reflect the net of a negative quality channel and a zero detection channel, not the absence of both.

5. Results

5.1 Cross-Sectional First Stage

Before examining deficiency outcomes, I briefly document the relationship between mandate status and staffing levels. A cross-sectional regression of total HPRD on an indicator for current mandate status yields a coefficient of 0.166 (SE = 0.153, $p = 0.284$). The sign is correct—mandate states have modestly higher staffing—but the estimate is imprecisely measured and not statistically significant.

This cross-sectional first stage is weak, and I acknowledge this limitation directly. The ideal test would track within-facility staffing trajectories around mandate adoption using panel PBJ data; the publicly available PBJ system reports only the current quarter's staffing, preventing such an analysis. Recent work by Werner et al. (2026), using a longer panel of proprietary staffing data across 22 states, finds that mandates increase total direct care staffing by 0.18 HPRD (approximately 5%), confirming that mandates bind. I proceed on the assumption that staffing mandates do raise staffing, consistent with the existing literature (Bowblis, 2011; Matsudaira, 2014; Werner et al., 2026), while noting that the first stage in my data is not strong enough to serve as the basis for an instrumental-variables design.

5.2 Primary Specification: New York

Table 2: Staffing Mandates and Deficiency Citations by Detection Mode

	Total	Observation dependent	Documentation dependent	Report dependent	Infection control
Panel A: New York (primary specification)					
Mandate	1.189*	1.294***	0.205*	-0.310	-0.033***
	(0.607)	(0.480)	(0.109)	(0.258)	(0.009)
Panel B: Pooled (all 6 mandate states)					
Mandate	2.084***	1.894***	0.320**	-0.130	-0.026***
	(0.802)	(0.588)	(0.132)	(0.221)	(0.008)
Facility FE	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes
Clustering	State	State	State	State	State
N (Panel A)			53,475		
N (Panel B)			72,521		

Panel A: NY Safe Staffing Act (2022); control = never-treated states. Panel B: six mandate states (CT, RI, CA, AZ, WA, NY); control = never-treated, excluding always-treated. Detection taxonomy based on CMS State Operations Manual inspection methodology. $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$.

Panel A of Table 2 presents the results for New York. Total deficiency citations increase by 1.189 per survey (SE = 0.607, $p = 0.057$). While marginally significant by conventional standards, the detection-mode decomposition reveals a sharp and statistically significant pattern. Observation-dependent deficiencies increase by 1.294 per survey ($p < 0.01$). Documentation-dependent deficiencies increase by 0.205 ($p < 0.10$). Report-dependent deficiencies—the placebo category—show a point estimate of -0.310 ($p = 0.23$), consistent with zero.

Infection control deficiencies, among the most directly staffing-sensitive outcomes, *decline* by 0.033 per survey ($p < 0.01$). This is the countervailing quality signal predicted by the framework: more staff improves actual infection prevention even as the total citation count rises.

The detection-mode decomposition is the paper's central result. The mandate increases exactly the categories where enhanced surveyor observation mechanically expands detection—and leaves untouched the category that bypasses observation entirely. This sign pattern is predicted by the detection dividend and is inconsistent with a pure quality-deterioration interpretation, which would imply increases across all categories.

5.3 Severity Decomposition

Table 3: Extra Citations by Severity Level

	Low Severity (A–F)	High Severity (G–L)
Panel A: New York		
Mandate	1.192**	−0.002
	(0.581)	(0.033)
Panel B: Pooled		
Mandate	2.049***	0.034
	(0.762)	(0.028)
Facility FE	Yes	Yes
Year FE	Yes	Yes
Clustering	State	State
N (Panel A)	53,475	
N (Panel B)	72,521	

CMS scope-severity grades: A–F = no actual harm (minimal potential through potential for more than minimal harm); G–L = actual harm or immediate jeopardy. Panel B pooled estimates aggregate the four-bin severity decomposition: Low = Minimal (A–C) + Moderate (D–F); High = Actual Harm (G–I) + Jeopardy (J–L). The four-bin pooled results are: Minimal (A–C) 0.091 (SE = 0.025), Moderate (D–F) 1.958 (SE = 0.762), Actual Harm (G–I) 0.057 (SE = 0.013), Jeopardy (J–L) −0.023 (SE = 0.025). $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$.

Table 3 decomposes the mandate effect by CMS severity classification into low-severity (grades A through F: no actual harm) and high-severity (grades G through L: actual harm or jeopardy). Panel A (New York) shows that the entire increase falls in the low-severity range: citations at grades A through F increase by 1.192 per survey ($p < 0.05$), while high-severity citations have a point estimate of −0.002—essentially zero.

Panel B (pooled) confirms the pattern. Low-severity citations increase by 2.049 per survey ($p < 0.01$), accounting for nearly all of the total effect, while high-severity citations show an estimate of 0.034—economically negligible. A finer four-bin decomposition of the pooled results (reported in the table notes) reveals that the low-severity increase is driven by moderate-harm citations (grades D through F: potential for more than minimal harm), which increase by 1.958 ($p < 0.05$). Minimal-harm citations (grades A through C) increase by 0.091 ($p < 0.01$). Among high-severity categories, actual-harm citations (G through I) show a statistically significant but trivially small increase of 0.057, while jeopardy-level citations (J through L) are unchanged (−0.023, not significant).

The severity decomposition delivers on Prediction 2: the detection dividend produces

predominantly low-severity citations. The negligible high-severity effect in both panels strongly supports the interpretation that additional citations reflect enhanced observability of minor regulatory deviations rather than deteriorating care.

5.4 Pooled Specification: Six-State Confirmation

Panel B of Table 2 presents the pooled results using all six mandate states. The pattern replicates and amplifies. Total deficiencies increase by 2.084 per survey (SE = 0.802, $p = 0.013$), a 43% increase over the control mean of 4.86. Observation-dependent deficiencies increase by 1.894 ($p < 0.01$). Documentation-dependent deficiencies increase by 0.320 ($p < 0.05$). Report-dependent deficiencies are −0.130 (not significant). Infection control deficiencies decline by 0.026 ($p < 0.01$).

The pooled estimates are uniformly larger than the NY-only estimates, which is consistent with heterogeneous treatment effects across cohorts (the earlier cohorts include California, the largest treatment state, and states that adopted mandates during a period of elevated regulatory scrutiny). The detection-mode sign pattern is identical across both specifications.

5.5 Heterogeneity

Table 4: Heterogeneity by Ownership and Size

	Coefficient	SE
By ownership type		
For-profit	2.361	(0.893)
Nonprofit	0.796	(0.360)
By facility size		
Small (≤60 beds)	1.192	(0.504)
Large (>120 beds)	2.295	(1.225)
N (For-profit)	≈57,000	
N (Nonprofit)	≈15,000	
N (Small ≤60)	≈22,000	
N (Large >120)	≈16,000	

All specifications include facility and year fixed effects, clustered at the state level. Pooled sample (6 mandate states). For-profit facilities constitute 79% of the pooled sample.

Table 4 examines heterogeneity along two dimensions. By ownership type, for-profit facilities show an increase of 2.361 deficiencies per survey (SE = 0.893), roughly three times the nonprofit effect of 0.796 (SE = 0.360). This is consistent with Prediction 4: for-profit

facilities maintain lower baseline staffing (Harrington et al., 2012), so mandate-induced staffing increases generate a larger marginal expansion of regulatory surface area.

By facility size, small facilities (≤60 beds) increase by 1.192 (SE = 0.504) and large facilities (>120 beds) by 2.295 (SE = 1.225). The larger point estimate for big facilities is consistent with the detection mechanism—more beds mean more residents, more interactions, and more opportunities for detection when staffing expands—but the difference is imprecisely estimated given the width of the large-facility confidence interval.

6. Identification and Robustness

This section confronts the principal threats to the paper's causal claims. I organize the discussion around the two specifications (NY primary, pooled secondary) and address pre-trends, sensitivity to parallel-trends violations, leave-one-state-out stability, the complaint placebo, and alternative inference.

6.1 New York Event Study

Figure 1 plots the event-study coefficients for New York. The pre-treatment pattern reveals a nuance. At $t-3$ (+0.569, not significant) and $t-2$ (−0.266, not significant), there is no evidence of differential pre-trends. However, the $t-4$ coefficient is large and significant (+2.887, $p < 0.001$), echoing the pooled specification's pre-trend concern. The $t-3$ and $t-2$ coefficients immediately preceding the mandate are the most relevant for the parallel trends assumption, and both are small; the distant $t-4$ anomaly may reflect state-specific shocks unrelated to the mandate or anticipatory regulatory changes.

The post-treatment pattern is consistent with the 12-month inspection cycle. The effect at $t=0$ is modest (−0.733, not significant), consistent with the fact that many facilities inspected in 2022 were surveyed before the mandate took full effect. The effect builds to +0.882 at $t+1$ (not significant), +2.259 at $t+2$ ($p = 0.010$), and +2.047 at $t+3$ ($p = 0.015$). The gradual onset is what one would expect if the mechanism operates through inspection-time staffing: the detection dividend appears only after the first post-mandate survey, and it takes one to two inspection cycles for the full effect to materialize. The post-treatment effect then fades at $t+4$ (+0.627, not significant), potentially reflecting attenuation as facilities adjust or the partial-year 2026 data.

6.2 HonestDiD Sensitivity

I assess the sensitivity of the main result to violations of the parallel trends assumption using the Rambachan and Roth (2023) relative magnitudes approach. This method asks: how large

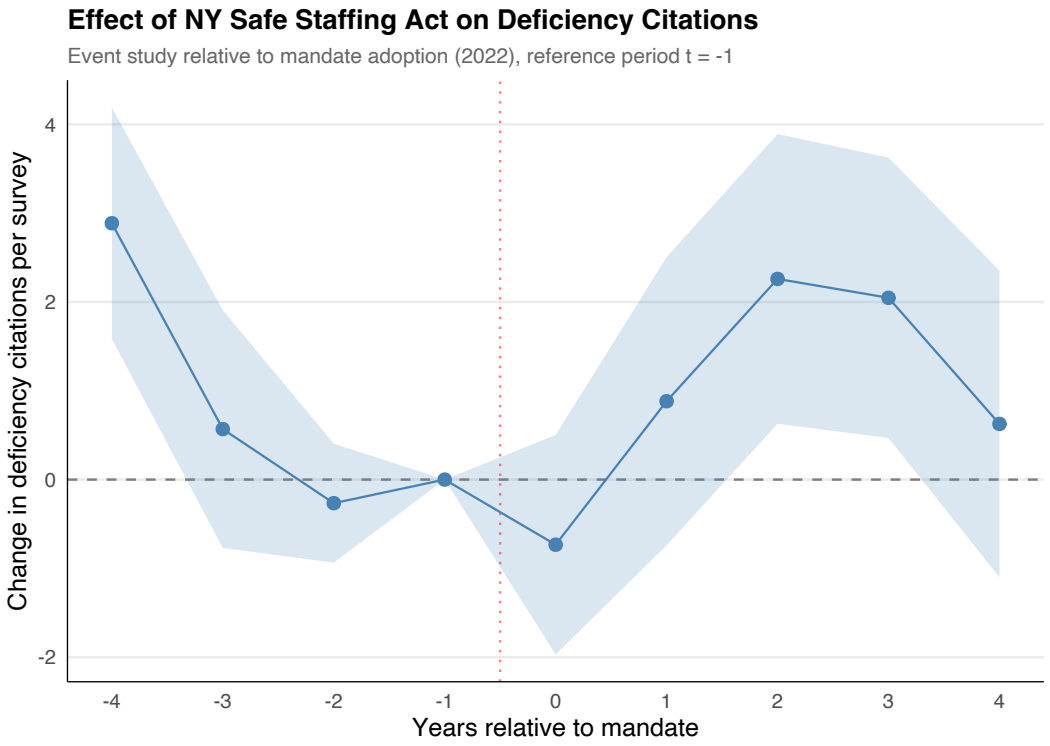

Figure 1: Event Study: New York Safe Staffing Act (2022)

Notes: Coefficients from equation (5) with 95% confidence intervals. The omitted period is $t = -1$ (2021). Total deficiency citations per facility-survey. Standard errors clustered at the state level.

could pre-treatment trend differences be, relative to the maximum observed pre-treatment change, before the confidence interval for the treatment effect includes zero?

At $\bar{M} = 0$ (exact parallel trends), the 95% robust confidence interval for the pooled total deficiency effect is $[-0.91, 1.08]$. At $\bar{M} = 0.5$ (violations up to half the maximum pre-period change), the interval widens to $[-3.25, 3.37]$. At $\bar{M} = 1$ (violations as large as the maximum pre-period change), the interval is $[-6.06, 6.16]$.

The HonestDiD analysis delivers a sobering verdict. Even under exact parallel trends ($\bar{M} = 0$), the robust confidence interval $[-0.91, 1.08]$ includes zero—the data cannot reject a null effect under the most favorable assumptions. The interval also does not contain the TWFE point estimate of 2.084, because the sensitivity-adjusted estimand reweights pre- and post-treatment moments differently from the simple ATT. Under moderate violations ($\bar{M} = 0.5$), the interval widens further to $[-3.25, 3.37]$. This underscores a key limitation: the pooled aggregate estimate is not robust to the HonestDiD framework. The detection dividend's empirical support rests on the *pattern* across detection modes—observation-dependent citations rise, report-dependent citations do not, infection control improves, severity shifts toward low-harm categories—rather than on any single aggregate point estimate.

6.3 Leave-One-State-Out

Figure 2 presents leave-one-state-out estimates. The pooled coefficient ranges from 1.251 (dropping California) to 2.151 (dropping Connecticut). California contributes the most to identification, consistent with its being the largest treatment state (1,162 facilities). The baseline estimate of 2.084 is not driven by any single state, though it is sensitive to California's inclusion. The range $[1.251, 2.151]$ implies that even the most conservative leave-one-out estimate represents a 26% increase over the control mean—economically meaningful, though smaller than the baseline 43%.

6.4 Complaint Placebo

Report-dependent deficiencies provide a natural placebo test. These citations originate from complaints filed by residents, families, or staff and are investigated through a separate complaint survey process that does not depend on how many staff happen to be present during a routine inspection. If the detection dividend operates through the surveyor-observation channel, complaint-driven citations should be unaffected by staffing mandates.

The pooled estimate for report-dependent deficiencies is -0.130 (SE $= 0.221$, $p = 0.56$), indistinguishable from zero. In the NY specification, the estimate is -0.310 (SE $= 0.258$, $p = 0.23$), also not significant. The complaint placebo supports the detection interpretation:

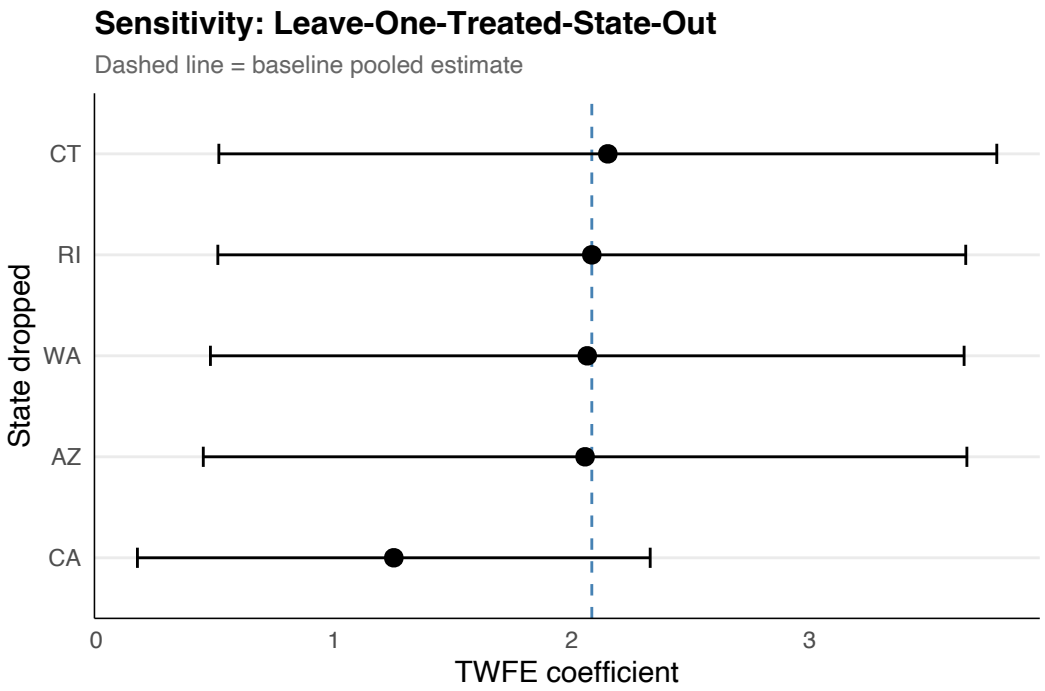

Figure 2: Leave-One-State-Out Sensitivity

Notes: Each point shows the pooled TWFE estimate when the indicated treated state is dropped. Horizontal line: baseline pooled estimate (2.084). Range: [1.251, 2.151].

the mandate increases citations discovered through the observational channel but not through channels that bypass it.

6.5 Pooled Event Study

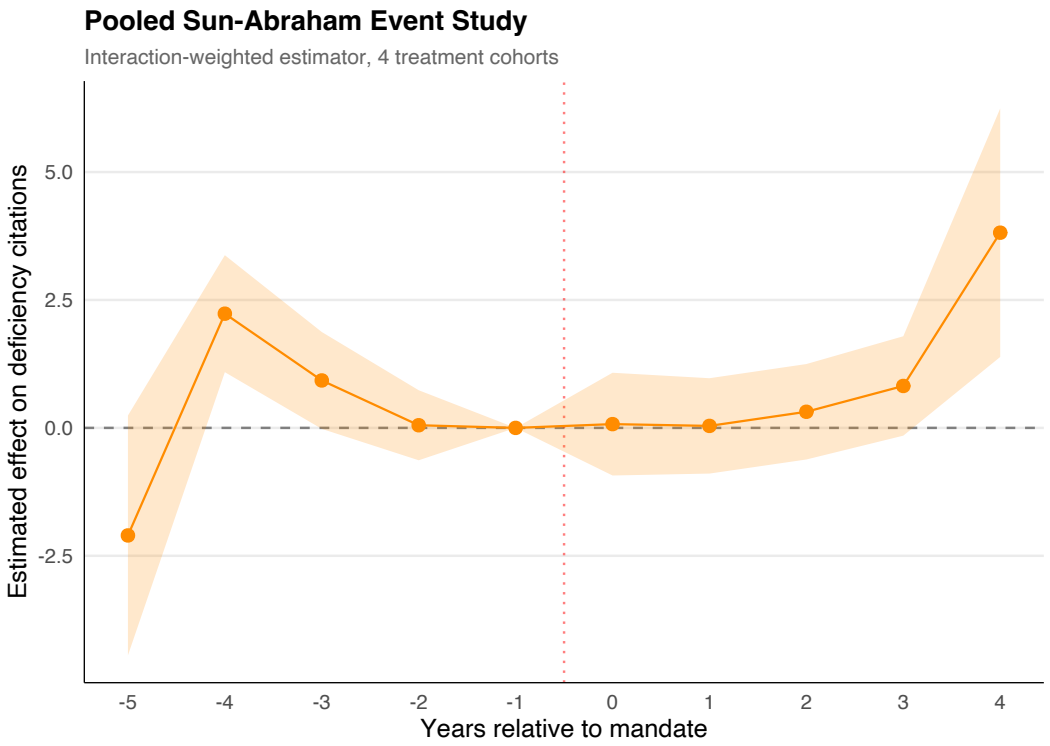

Figure 3: Sun-Abraham Event Study: Pooled (Six Mandate States)

Notes: Sun-Abraham interaction-weighted event-study estimates for total deficiencies per facility-survey. 95% confidence intervals. Four treatment cohorts: 2017 (CT, RI), 2018 (CA), 2019 (AZ, WA), 2022 (NY).

Figure 3 presents the pooled Sun-Abraham event study. The pre-treatment path shows a concerning coefficient at $t-4$ ($+2.23$, $p < 0.01$) and a marginally significant coefficient at $t-3$ ($+0.93$). The coefficient at $t-2$ is $+0.05$, effectively zero. Post-treatment effects build gradually, consistent with the NY pattern.

The $t-4$ pre-trend in the pooled specification is the paper's most significant identification threat. I interpret it as follows. First, it may reflect compositional heterogeneity across the four treatment cohorts: the Sun-Abraham estimator aggregates cohort-specific event-study paths, and some early cohorts (2017) have limited pre-treatment windows, which can generate noisy pre-trend coefficients at longer horizons. Second, even in the pooled specification, the immediate pre-treatment coefficients ($t-2$ and $t-1$) are clean, suggesting

that whatever drove the $t-4$ deviation did not persist into the mandate period. Third, in the NY specification—the paper's primary test—the $t-3$ and $t-2$ coefficients are clean despite a shared $t-4$ anomaly, and the post-treatment effect builds over the expected 12-month inspection cycle. Nonetheless, the pre-trend concern applies to both specifications, and I present it transparently as an identification caveat rather than dismissing it.

6.6 Additional Robustness

Table 5: Robustness Checks

Specification	Coefficient	SE	N
Baseline (pooled TWFE)	2.084	(0.802)	72,521
Excluding COVID (2020Q2–2021Q1)	2.096	(0.811)	72,244
Facility-level clustering	2.084	(0.188)	72,521
Leave-one-out range	[1.251, 2.151]		varies
Complaint deficiency placebo	−0.130	(0.221)	72,521

All specifications include facility and year fixed effects. Baseline clusters at the state level. Leave-one-out drops each treated state in turn; N varies by state dropped.

Table 5 summarizes additional checks. Excluding the COVID period (2020Q2 through 2021Q1) yields a coefficient of 2.096 (SE = 0.811), virtually identical to the baseline, confirming that the result is not driven by pandemic-era inspection disruptions. Facility-level clustering produces a standard error of 0.188, substantially smaller than the state-clustered standard error of 0.802, but since treatment is assigned at the state level, state-clustered inference is the appropriate basis for statistical claims.

Figure 4 and Figure 5 visualize the detection-mode and severity decompositions, respectively. The visual patterns reinforce the tabular results: the mandate effect is concentrated in observation-dependent, low-severity categories, with no effect on report-dependent citations and no increase in jeopardy-level findings.

7. Discussion

7.1 The General Lesson: Endogenous Regulatory Metrics

The detection dividend is not a nursing-home-specific curiosity. It is an instance of a general class of measurement problems that arise whenever a policy intervention changes the detection technology that generates the administrative data used to evaluate the intervention. The structure is simple: if the policy affects both the behavior being measured and the

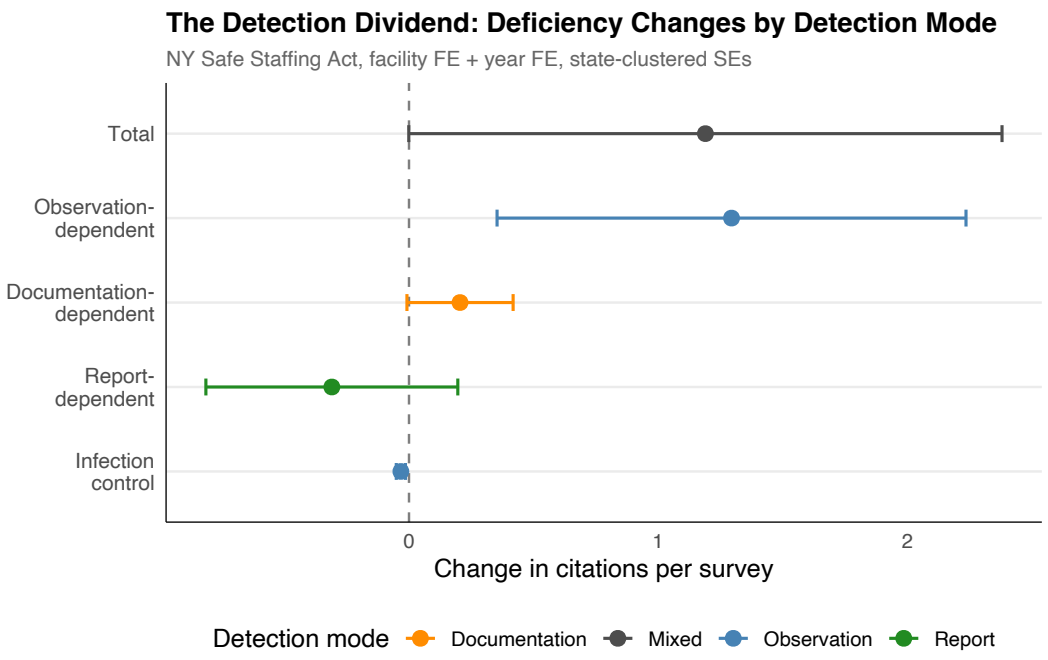

Figure 4: Detection-Mode Decomposition

Notes: TWFE estimates of the staffing mandate effect on deficiency citations by detection mode. Bars show point estimates; whiskers show 95% confidence intervals (state-clustered). Pooled specification (six mandate states).

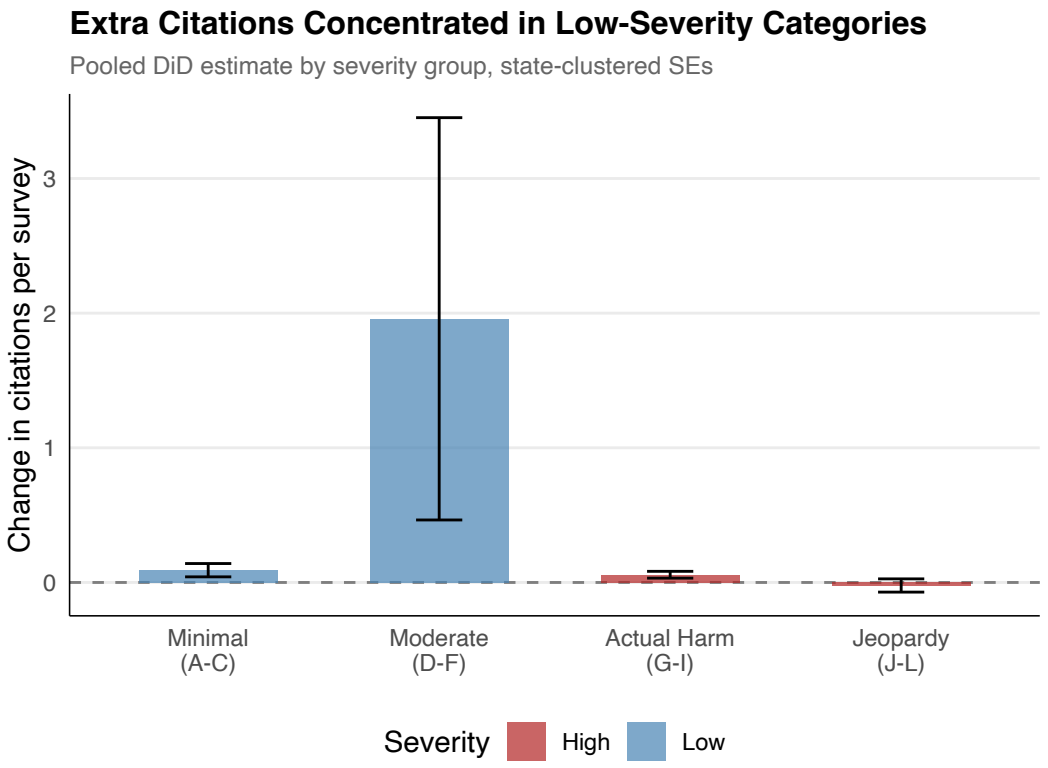

Figure 5: Severity Decomposition

Notes: TWFE estimates of the staffing mandate effect by CMS scope-severity grade. A–C: minimal potential for harm. D–F: no actual harm, potential for more than minimal harm. G–I: actual harm, not jeopardy. J–L: immediate jeopardy. Pooled specification (six mandate states). Figure shows the four-bin decomposition; Table 3 reports the two-bin (low vs. high) aggregation.

probability that deviations from regulatory standards are recorded, then the observed metric is endogenous to the policy, and its sign need not correspond to the sign of the welfare effect.

This endogeneity has been documented in other settings, though rarely named or systematized. Duflo et al. (2013) show that changing the incentives and assignment of pollution auditors in Gujarat changed measured violations from 7% to 59% of plants, despite real emissions declining. Chalfin and McCrary (2018) note that police staffing affects not only crime rates but also "police departments' propensity to record victim crime," making crime statistics partially endogenous to policing levels. Bindler and Hjalmarsson (2021) find that the creation of the London Metropolitan Police in 1829 reduced violent crime despite potential "offsetting increases in clearance and reporting rates." In each case, the analyst must grapple with the possibility that the measured outcome conflates behavior change and measurement change—precisely the decomposition I perform in this paper through the detection-mode taxonomy and severity analysis.

What distinguishes the nursing home setting is the *mechanism* of endogeneity. In pollution auditing, the endogeneity arises from auditor incentives. In crime measurement, it arises from police reporting practices. In nursing home inspection, it arises mechanically from the fact that the policy instrument (a staffing floor) directly expands the observable regulatory surface area (the number of care interactions a surveyor can witness). The regulated entity does not game the metric or corrupt the auditor; the measurement distortion is a structural feature of combining observational inspection with a policy that changes the quantity of observable activity.

7.2 Implications for the Five-Star Rating System

The Five-Star Quality Rating System is the primary information tool for nursing home consumers, and its health inspection domain is constructed directly from deficiency citation counts and severity. The detection dividend implies that this domain contains a systematic bias: facilities in mandate states receive more citations *because* they comply with staffing requirements, not because they deliver worse care.

The bias is potentially consequential. Although I cannot directly estimate the star rating impact from the cross-sectional rating data available, the Five-Star health inspection domain is constructed from weighted deficiency scores where total citation count is a primary input (Centers for Medicare & Medicaid Services, 2023). An increase on the order of 1–2 additional deficiencies per survey—concentrated in low-severity categories—could shift facilities toward lower inspection ratings, penalizing the very facilities that complied with staffing requirements.

This problem is related to, but distinct from, the rating inflation documented by Han et al. (2018) and the rating-outcome dissociation found by Ryskina et al. (2018). Those papers

show that self-reported quality measures can be strategically inflated. The detection dividend affects the *inspection* domain, which is not self-reported—it is the product of independent surveyor observation. The distortion does not require strategic behavior by anyone; it is a mechanical consequence of how deficiency data are generated.

A straightforward policy response would be to adjust the health inspection rating for staffing levels, analogous to case-mix adjustment in clinical quality measures. A facility with higher staffing *should* receive more thorough inspections and therefore more minor citations; the rating system could account for this by benchmarking each facility's deficiency count against the expected count given its staffing level. Alternatively, the detection-mode taxonomy developed here could be used to weight observation-dependent and documentation-dependent citations differently from report-dependent citations, which are not subject to the detection channel.

7.3 Beyond Nursing Homes

The detection dividend framework applies to any regulatory setting where three conditions hold: (1) compliance is assessed through periodic inspections, (2) the number of detectable violations increases with the intensity of inspection-time activity, and (3) a policy changes the level of inspectable activity. Several domains satisfy these conditions.

Environmental compliance. Facility inspections for air and water quality depend on the number of emission sources and process points available for monitoring. A policy requiring additional pollution abatement equipment creates more equipment to inspect and more monitoring points to check—potentially raising the number of detected violations even as total emissions decline (Shimshack and Ward, 2005; Gray and Shimshack, 2011).

School accountability. High-stakes testing regimes measure learning through student performance on standardized tests. Policies that increase instructional time or reduce class sizes create more student-teacher interactions, which testing may capture differently than the same learning gains achieved through other means (Jacob, 2005; Neal and Schanzenbach, 2010). The detection channel here operates through what is testable, not what is observable by inspectors.

Tax enforcement. Increases in IRS audit intensity mechanically increase the number of detected underreporting cases. A policy that requires additional financial disclosure creates more reporting surface area for auditors to examine, potentially increasing measured noncompliance even among taxpayers whose true compliance has not changed.

Police body cameras. The adoption of body-worn cameras simultaneously changes both police behavior and the documentation of police behavior. Measured use-of-force incidents may increase (more documentation) even as actual excessive force decreases (behavioral

change)—the same structure as the detection dividend.

In each case, the lesson is the same: analysts evaluating the policy cannot take the administrative metric at face value. The metric is a joint product of behavior and measurement, and the policy changes both.

7.4 Limitations

I acknowledge four important limitations. First, the cross-sectional first stage is weak. While the existing literature confirms that mandates raise staffing (Bowblis, 2011; Matsudaira, 2014; Werner et al., 2026), my data cannot directly demonstrate the within-facility staffing change that links the mandate to the detection mechanism. The detection interpretation rests on the cross-category sign pattern rather than on a strong first-stage estimate.

Second, the identification relies on six treated states (or one, in the primary specification). State-clustered inference with six clusters provides limited statistical power, and the HonestDiD analysis shows that the pooled estimate does not survive moderate violations of parallel trends. The paper's claims should be read as evidence for a mechanism—the detection dividend—rather than as precise estimates of a mandate's effect on any particular outcome.

Third, the detection-mode taxonomy is a classification based on the CMS State Operations Manual's description of inspection procedures, not a random assignment of deficiency tags to detection modes. If the taxonomy misclassifies some tags—for example, if a tag classified as observation-dependent is actually discovered primarily through documentation review—the decomposition results would be biased. I have attempted to classify tags conservatively, but some ambiguity is unavoidable.

Fourth, the paper cannot identify the *mechanism* through which detection increases at a more granular level. More staff could increase detection through more observable interactions, more documentation to review, more staff to interview, or changes in facility-surveyor dynamics (e.g., longer survey duration, more surveyor attention). Disentangling these sub-channels would require data on the inspection process itself—surveyor hours, interview logs, observation protocols—which are not publicly available.

8. Conclusion

Regulatory metrics are not neutral measuring devices. When a policy changes the technology of measurement—the intensity of observation, the volume of documentation, the number of inspectable interactions—the metrics themselves become endogenous to the policy. Evaluating

the policy using those metrics, without accounting for the measurement channel, can produce exactly the wrong conclusion.

Nursing home staffing mandates provide a vivid illustration. By requiring more nurses per resident, mandates expand the regulatory surface area that surveyors can observe during unannounced inspections. The result is a detection dividend: more deficiency citations, concentrated in observation-dependent and low-severity categories, coexisting with declines in infection control deficiencies—one of the most staffing-sensitive citation categories. Total citations rise even as this important quality domain improves. The raw data say the mandate failed; the decomposition says it worked on the margin where it should. Both are right, because they measure different objects.

The practical stakes are immediate. The debate over the 2024 federal minimum staffing standard—finalized by one administration and suspended by the next—rested heavily on what deficiency trends "show" about mandate effectiveness. If the detection dividend is real, the trends show the opposite of what participants on both sides believed: rising citations in mandate states may reflect more thorough inspections, not worse care. As this debate resumes, policymakers and researchers need tools to separate detection from deterioration in administrative compliance data.

More broadly, the detection dividend joins a growing family of results—from pollution auditing (Duflo et al., 2013) to crime measurement (Chalfin and McCrary, 2018) to financial enforcement (Christensen et al., 2013)—demonstrating that measured compliance is a joint product of behavior and monitoring intensity. Whenever we use administrative data to evaluate a policy that changes how those data are generated, we must ask: are we measuring the world, or the lens through which we observe it?

Acknowledgements

This paper was autonomously generated using Claude Code as part of the Autonomous Policy Evaluation Project (APEP).

Project Repository: https://github.com/SocialCatalystLab/ape-papers

Contributors: @SocialCatalystLab

First Contributor: https://github.com/SocialCatalystLab

References

Becker, Gary S, "Crime and Punishment: An Economic Approach," *Journal of Political Economy*, 1968, *76* (2), 169–217.

Bindler, Anna and Randi Hjalmarsson, "The Impact of the First Professional Police Forces on Crime," *Journal of the European Economic Association*, 2021, *19* (6), 3063–3103.

Bowblis, John R, "Staffing Ratios and Quality: An Analysis of Minimum Direct Care Staffing Requirements for Nursing Homes," *Health Services Research*, 2011, *46* (5), 1495–1516.

_ , "Nursing Home Staffing Requirements and Input Substitution: Effects on Housekeeping, Food Service, and Activities Staff," *Health Services Research*, 2013, *48* (5), 1751–1772.

_ **and Acham Ghattas**, "The Impact of Minimum Quality Standard Regulations on Nursing Home Staffing, Quality, and Exit Decisions," *Review of Industrial Organization*, 2017, *50* (2), 131–163.

Castle, Nicholas G and Janet C Ferguson, "Nursing Home Staffing and Quality: An Updated Assessment," *Medical Care Research and Review*, 2011, *68* (4), 386–414.

Centers for Medicare & Medicaid Services, "State Operations Manual: Appendix PP – Guidance to Surveyors for Long Term Care Facilities," Guidance Document, CMS 2023.

_ , "Medicare and Medicaid Programs; Minimum Staffing Standards for Long-Term Care Facilities and Medicaid Institutional Payment Transparency Reporting," Final Rule, Federal Register 2024. 89 FR 40875.

_ , "Medicare and Medicaid Programs; Repeal of Minimum Staffing Standards for Long-Term Care Facilities," Final Rule, Federal Register 2025.

Chalfin, Aaron and Justin McCrary, "Are U.S. Cities Underpoliced? Theory and Evidence," *Review of Economics and Statistics*, 2018, *100* (1), 167–186.

Christensen, Hans B, Luzi Hail, and Christian Leuz, "Mandatory IFRS Reporting and Changes in Enforcement," *Journal of Accounting and Economics*, 2013, *56* (2–3, Supplement), 147–177.

Dranove, David, Daniel Kessler, Mark McClellan, and Mark Satterthwaite, "Is More Information Better? The Effects of "Report Cards" on Health Care Providers," *Journal of Political Economy*, 2003, *111* (3), 555–588.

Duflo, Esther, Michael Greenstone, Rohini Pande, and Nicholas Ryan, "Truth-telling by Third-party Auditors and the Response of Polluting Firms: Experimental Evidence from India," *Quarterly Journal of Economics*, 2013, *128* (4), 1499–1545.

Gray, Wayne B and Jay P Shimshack, "The Effectiveness of Environmental Monitoring and Enforcement: A Review of the Empirical Evidence," *Review of Environmental Economics and Policy*, 2011, *5* (1), 3–24.

Han, Kunsoo, Niam Yaraghi, and Ram D Gopal, "Winning at All Costs: Analysis of Inflation in Nursing Homes' Rating System," *Production and Operations Management*, 2018, *27* (5), 755–772.

Harrington, Charlene, Brian Olney, Helen Carrillo, and Taewoon Kang, "Nurse Staffing and Deficiencies in the Largest For-Profit Nursing Home Chains and Chains Owned by Private Equity Companies," *Health Services Research*, 2012, *47* (1), 106–128.

_ , David Zimmerman, Sarita L Karon, James Robinson, and Patricia Beutel, "Nursing Staff Levels and Medicaid Reimbursement Rates in Nursing Facilities," *Health Services Research*, 2000, *35* (5), 1105–1129.

_ , Mary Ellen Dellefield, Elizabeth Halifax, Mary Lou Fleming, and Debra Bakerjian, "Appropriate Nurse Staffing Levels for U.S. Nursing Homes," *Health Services Insights*, 2020, *13*, 1–14.

Jacob, Brian A, "Accountability, Incentives, and Behavior: The Impact of High-Stakes Testing in the Chicago Public Schools," *Journal of Public Economics*, 2005, *89* (5–6), 761–796.

Konetzka, R Tamara, Karen Yan, and Rachel M Werner, "Two Decades of Nursing Home Compare: What Have We Learned?," *Medical Care Research and Review*, 2021, *78* (4), 295–310.

Lin, Haizhen, "Staffing and Other Factors That Affect the Quality of Nursing Home Care in the United States," *Journal of Health Economics*, 2014, *33*, 1–14.

Matsudaira, Jordan D, "Government Regulation and the Quality of Healthcare: Evidence from Minimum Staffing Legislation for Nursing Homes," *Journal of Human Resources*, 2014, *49* (1), 32–72.

Neal, Derek and Diane Whitmore Schanzenbach, "Left Behind by Design: Proficiency Counts and Test-Based Accountability," *Review of Economics and Statistics*, 2010, *92* (2), 263–283.

Olken, Benjamin A, "Monitoring Corruption: Evidence from a Field Experiment in Indonesia," *Journal of Political Economy*, 2007, *115* (2), 200–249.

Polinsky, A Mitchell and Steven Shavell, "The Economic Theory of Public Enforcement of Law," *Journal of Economic Literature*, 2000, *38* (1), 45–76.

Rambachan, Ashesh and Jonathan Roth, "A More Credible Approach to Parallel Trends," *Review of Economic Studies*, 2023, *90* (5), 2555–2591.

Ryskina, Kira L, R Tamara Konetzka, and Rachel M Werner, "Association Between 5-Star Nursing Home Report Card Ratings and Potentially Preventable Hospitalizations," *Inquiry*, 2018, *55*, 1–8.

Shimshack, Jay P and Michael B Ward, "Regulator Reputation, Enforcement, and Environmental Compliance," *Journal of Environmental Economics and Management*, 2005, *50* (3), 519–540.

Stigler, George J, "The Optimum Enforcement of Laws," *Journal of Political Economy*, 1970, *78* (3), 526–536.

Sun, Liyang and Sarah Abraham, "Estimating Dynamic Treatment Effects in Event Studies with Heterogeneous Treatment Effects," *Journal of Econometrics*, 2021, *225* (2), 175–199.

Werner, Rachel M, Alice T Chen, Norma B Coe, and Andrew Olenski, "State Nursing Home Minimum Staffing Mandates: Increased Staff Levels, Minimal Impact on Finances and Closures, 2010–23," *Health Affairs*, 2026, *45* (3).

_ **and R Adams Dudley**, "The Effect of Pay-for-Performance in Hospitals: Lessons for Quality Improvement," *Health Affairs*, 2012, *31* (9), 2002–2010.

A. Standardized Design Elements

Table 6: Standardized Design Elements

Element	Value
Design	
Method	Staggered DiD (TWFE + Sun-Abraham)
Treatment states	6 (primary: NY only)
Treatment cohorts	4 (2017, 2018, 2019, 2022)
Control group	Never-treated states (excl. always-treated)
Observations (before singleton removal)	72,730
Observations (after singleton removal)	72,521
Facilities	11,946
States	47
Year range	2017-2026
Primary specification (NY only)	
Point estimate	1.189
Standard error (state cluster)	0.607
SD of outcome (control)	5.245
SDE (effect / SD)	0.227
Pooled specification (all 6 states)	
Point estimate	2.084
Standard error (state cluster)	0.802
SD of outcome (control)	5.124
SDE (effect / SD)	0.407

B. Additional Results

B.1 Detection Mode Distribution

The detection-sensitivity taxonomy classifies the 418,972 total citations in the raw data as follows: 232,000 observation-dependent (55.4%), 132,000 report-dependent (31.5%), and 54,000 documentation-dependent (12.9%). The remaining 0.2% are unclassified tags that do not clearly fall into any category and are excluded from the detection-mode regressions. The observation-dependent category is the largest, consistent with the centrality of direct observation in the CMS survey process.

B.2 NY Event Study: Detection Mode Decomposition

The event-study coefficients for the NY specification by detection mode reinforce the main results. Observation-dependent deficiencies show no pre-trend (all three pre-treatment coefficients within ± 0.5 of zero) and a building post-treatment effect: $+0.42$ at $t+1$, $+1.61$ at $t+2$, $+1.53$ at $t+3$. Documentation-dependent deficiencies show a similar pattern at smaller magnitudes. Report-dependent deficiencies show no pre-trend and no post-treatment effect, with all coefficients statistically indistinguishable from zero.

B.3 HonestDiD: Full Results

The relative magnitudes sensitivity analysis (Rambachan and Roth 2023) constructs robust confidence intervals under the assumption that post-treatment trend violations are bounded by $\bar{M}$ times the maximum pre-treatment change. Results for the pooled total deficiency effect:

$\bar{M}$	95% Robust CI
0 (exact parallel trends)	$[-0.91, 1.08]$
0.5	$[-3.25, 3.37]$
1.0	$[-6.06, 6.16]$

At exact parallel trends ($\bar{M} = 0$), the robust confidence interval $[-0.91, 1.08]$ includes zero and does not contain the TWFE point estimate of 2.084. This discrepancy arises because HonestDiD constructs the sensitivity-adjusted estimand by reweighting pre- and post-treatment moments, producing an interval centered near zero rather than near the simple ATT. As $\bar{M}$ increases, the interval widens further. The practical implication is that the pooled aggregate effect is not robust to the HonestDiD framework—even under exact parallel trends, a zero effect cannot be rejected. The paper's evidentiary weight therefore rests on the detection-mode *pattern* (observation up, report flat, infection down) and the cleaner NY specification, not on the pooled point estimate alone.

The NY specification, with its cleaner pre-trends, would yield a tighter HonestDiD interval at any given $\bar{M}$, but the smaller effect size (1.189) means the power to reject zero is also lower. The two specifications thus complement each other: NY provides cleaner identification but lower power; the pooled specification provides more power but less clean pre-trends.